AfterWork

Market Basket Analysis

Learning Outcomes

By the end of this topic, you will have achieved the following learning outcomes:

- Explain what association rules and item sets are in market basket analysis.
- Describe the basic process for performing market basket analysis.
- Find association rules to solve data science problems using the Python programming language.

"Predicting the future isn't magic. It's artificial intelligence." ~Dave Waters

What is Market Basket Analysis

Market basket analysis is a type of analysis that identifies the strength of association between pairs of products purchased together and identifies patterns of co-occurrence. A co-occurrence is when two or more things take place together.

In other words, market basket analysis describes how often the items appear together in historical transactions.

In the process, market basket analysis creates If-Then scenario rules (association rules) derived from the frequencies of co-occurrence in the observations. For example, if item A is purchased, then item B is likely to be purchased.

Market basket analysis is also known as association analysis or association rule mining.

Example

The most commonly cited example of market basket analysis is the "beer and diapers" case, where a large retailer performed market basket analysis on their transaction data and found an unexpected purchase pattern of individuals buying beer and baby diapers at the same time.

Why Market Basket Analysis?

Across organizations, market basket analysis provides the potential for:

- Increasing customer engagement
- Boosting sales and increasing RoI
- Improving customer experience
- Optimizing marketing strategies and campaigns
- Helping to understand customers better
- Identifying customer behavior and pattern

Questions that we can answer using market basket analysis include:

- What products are purchased together most frequently?
- How should the products be organized and positioned in the store?
- How do we identify the best products to discount via coupons?

Applications

Market basket analysis is widely used across various sectors, i.e., retail, banking, telecommunication, etc. The discovered associations help stakeholders develop marketing strategies such as optimizing store layouts, design of product bundles, discount/coupon offers, etc.

Retail

- Market Basket Analysis can help determine what items are purchased together, purchased sequentially, and purchased by season.
- Outcomes include:
  - Designing store layout so that consumers can more easily find items that are frequently purchased together.
  - Recommending associated products that are frequently bought together, “Customers who purchased this product also viewed this product…”. Does it make sense to sell soda and chips or soda and crackers?
  - Emailing customers who bought specific products with other products and offered those products that are likely to be attractive to them.
  - Grouping products that customers frequently purchase together in the store’s product placement
  - Designing special promotions that combine or discount certain products
  - Optimizing the layout of the catalog of an eCommerce site
  - Controlling inventory based on product demands and what products sell better together

Telecommunications

- In Telecommunications, where high churn rates continue to be a growing concern, we can use market Basket Analysis to determine what services are being utilized and what packages customers are purchasing. They can use that knowledge to direct marketing efforts at customers who are more likely to follow the same path.

    - For instance, Telecommunications these days is also offering TV and Internet. Creating bundles for purchases can be determined by analyzing what customers purchase, thereby giving the company an idea of how to price the bundles. This analysis might also lead to determining the capacity requirements.

Banks

- In financial (banking, for instance), we can use market Basket Analysis to analyze customers' credit card purchases to build profiles for fraud detection purposes and cross-selling opportunities.

Insurance

- In insurance, we can use market Basket Analysis to build profiles to detect medical insurance claim fraud. By building claims profiles, you can then use the profiles to determine if more than one claim belongs to a particular claim within a specified time.

Medicine

- In healthcare or medicine, we can use market basket analysis for symptom analysis, with which we can better identify a profile of illness. We can also use it to reveal biologically relevant associations between genes or environmental effects and gene expression.

Association Rules

Association rules are typically written like this: {Diapers} -> {Beer} which means that there is a strong relationship between customers that purchased diapers and also purchased beer in the same transaction. These rules highlight frequent patterns of associations among sets of items or objects in transaction databases.

Key Association Rule Terminologies

- In the above beer-diaper example, the {Diaper} is the **antecedent,** and the {Beer} is the **consequent**. Both antecedents and consequents can have multiple items. In other words, {Diaper, Gum} -> {Beer, Chips} is a valid associative rule.
- An **itemset** is a collection of one or more items e.g. {Milk, Bread, Diaper}
- An **k-itemset** is a set of k items e.g. {beer, diapers, juice} is a 3-itemset; {cheese} is a 1-itemset; {honey, ice-cream} is a 2-itemset.
- A **transaction** is a single customer purchase, and the items are the things that were bought.
- An **association rule** is a statement of the form {item set A} -> {item set B}.

Support, **Confidence**, and **Lift** are three important evaluation metrics for finding association rules.

Support

- This is a metric that indicates how frequently the itemset occurs within the dataset. It gives the fraction of transactions that contain items x and y.

  $$\text{Support}(A => B) = P(A \cap B)$$

- For example, if your itemset {beer, diapers} appears in 10% of transactions in the whole dataset, it will have a support of 0.1. Or take five transactions, for instance. If you purchase bread in 3 transactions, you can tell the support of bread is equal to 3/5.
- In many instances, you may want to look for high support to ensure it is a good relationship. However, there may be instances where low support is helpful if you are trying to find “hidden” relationships.

Confidence

- This is a metric that gives us the likelihood of certain items being purchased together. It tells us how often the items x and y occur together, given the number of times x occurs.

  $$\text{Confidence}(A => B) = P(B \mid A)$$

- For example, how likely is diapers purchased when beer is purchased, i.e., the proportion of transactions containing diapers containing beer.
- Confidence values range from 0 to 1, where 0 indicates that Y is never purchased when X is purchased, and 1 indicates that Y is always purchased whenever X is purchased. A value of 1 indicates that the itemset occurs 100% of the time, while 0.1 shows that it occurs 10% of the time. A 50% confidence may be perfectly acceptable for product recommendations, but this level may not be high enough in a medical situation.

Lift

- This metric gives us the probability of all of the items in a rule occurring together (otherwise known as the support) divided by the product of the probabilities of the items on the left and right-hand side occurring as if there was no association between them.

  $$Lift(A \Rightarrow B) = \frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

- For example, if beer, diapers, and chips occurred together in 2.5% of all transactions, beer and diapers in 10% of transactions, and chips in 8% of transactions, the lift would be 0.025/(0.1*0.08) = 3.125. In such a case, it says how likely beer is purchased when a diaper is purchased while controlling for how popular beer is.

- **Lift = 1:** This implies no relationship between beer and diapers (i.e., beer and diapers occur together only by chance). The rules were utterly independent. Thus, no inference can be made about beer when the diaper is purchased.
- **Lift > 1:** This implies a positive relationship between beer and diapers (i.e., beer and diapers occur together more often than randomly). It means that beer is likely to be purchased together with diapers. When we perform market basket analysis, we are looking for rules with a lift of more than 1.
- **Lift < 1:** This implies that there is a negative relationship between beer and diaper (i.e., beer and diaper occur together less often than random)

Interpretation Examples

1. **Example 1**
   - **Association Rule:** {bread} -> {butter}
   - **Metrics:** Support = 0.36, Confidence = 1.0, Lift = 1.83
   - **Interpretation:** 36% of transactions contain both bread and butter. Butter appears every time in transactions that contain bread only. Confidence = 1 indicates that butter is always purchased whenever bread is purchased. Lastly, the value of lift is greater than 1, and it means it is more likely bread and butter will be bought together than each individually.
2. **Example 2**
   - **Association Rule:** {P1} -> {P2}
   - **Metrics:** Support: 0.240, Confidence: 0.750, Lift: 1.923
   - **Interpretation:** If a customer bought product P1, there is a 75% chance that they will buy product 2.
3. **Example 3**
   - **Association Rule:** {P10, P31} -> {P64}
   - **Metrics:** Confidence: 0.810, Suppose: 0.170, Lift: 2.188
   - **Interpretation:** If a customer bought the products 10 and 31, there is an 81% chance that they will buy product 64.

Market basket analysts search for rules with a lift greater than 1 backed with high confidence values and often high support.

Rule Mining Algorithms

We use rule mining algorithms to search for associative rules within a dataset. The **Apriori Algorithm** is such a popular algorithm.

This algorithm helps mine frequent itemsets and relevant association rules. It works on a database containing a large number of transactions. A frequent itemset is an itemset whose support is greater than or equal to a minimum threshold (that we set).

Apriori Algorithm Steps

The steps of working of the apriori algorithm can be given as:

- Transform the data to a 1-hot encoded data frame.
- Use the rule mining algorithm to generate the frequent itemsets with higher support(sup) than the minimum support.
    - You also define the minimum support
- Generate the rules with their corresponding support, confidence, and lift.
    - You also define lift
- Sort these association rules in decreasing order.
- Analyze the rules along with their confidence and support.

We can also use other association rule algorithms such as Eclat and FP-Growth. We'll leave those algorithms for your further learning.

Tools

- We can use Python to find associative rules. Python libraries that would come in handy in this process include would be NumPy, Pandas, and Mlxetend libraries.

NB: It's important to note that association rules do not extract an individual's preference, instead find relationships between sets of elements of every particular transaction.

References

You can also use the following resources for further reading.

1. Introduction to Market Basket Analysis [Link]
2. How to perform Market Basket Analysis [Link]
3. Basics Association Analysis [Link]
4. Comprehensive Guide to Market Basket Analysis [Link]
5. Affinity Analysis: Market Basket Analysis [Link]